

PROJECT 1 OVERVIEW

Reengineer DOT Data Programs

BACKGROUND

In September 2000, the Bureau of Transportation Statistics (BTS) published the Safety Data Action Plan with the goal of providing the U.S. Department of Transportation (DOT) with quality data, capable of identifying, quantifying, and minimizing risk factors in U.S. travel. The Safety Data Action Plan identified 10 research projects to address specific shortcomings in current data collection and data quality within the various DOT databases. The first research project addresses reengineering DOT data programs.

DOT maintains in excess of 40 programs that capture either safety data or crucial related information, such as measures of exposure. A data quality review requested by Congress indicated that quality improvements can be made that will better serve the DOT mission. It was decided that the first step in reengineering data programs is a data quality audit of all major safety data systems to evaluate existing capabilities and determine needed improvements. This includes review and assessment of DOT data collection systems, as well as other transportation safety data systems not directly collected by DOT, but accessible within DOT data systems. The next step would be to implement recommendations for improvements, based on the assessment performed. This overview will focus on the data quality assessments phase. With improved data, DOT's safety programs will become not only more effective, but more cost-effective as well. The

Department can better address its strategic goal of improving safety by developing more targeted inspection, education, regulatory, and research programs.

Objective

As previously mentioned, the initial goal of this project is to conduct quality audits of transportation safety data systems. Due to resource constraints, BTS decided to conduct data quality assessments of five major data systems by the end of 2001. Data quality is a broad concept that refers, ultimately, to the usefulness of data for analysis and decisionmaking. The overall objective of this project is to ensure that decisionmakers can have a reasonable level of confidence in the source and reliability of transportation safety data.

Process

A data quality assessment template was developed to guide the person responsible for the assessment and to afford consistency between assessments. The template includes the following sections: Background, Frames and Sampling, Data Collection, Data Preparation, Data Dissemination, Sponsor Evaluation, Data Analysis, Assessment, and Recommendations and Suggestions for Data Quality Improvements. See the Attachment for details.

The selection of data systems was based on recommendations from the Safety Data Task Force members and include: the UNISHIP data system, the Hazardous Materials

Management Information System (HMIS), the Airline Passenger Origin and Destination Survey, the National Transit Database System – Safety & Security module, and the National Aviation Safety Data Analysis Center (NASDAC)) data system. Draft Data Quality Assessment Reports for each of the five data systems are under management review. Both the assessment and the recommendations for each system aim at improving the relevance, completeness, quality, timeliness, comparability and utility of transportation safety data.

DATA SYSTEM SPECIFICS AND RECOMMENDATIONS FOR DATA QUALITY IMPROVEMENTS

UNISHIP

UNISHIP is an enforcement database for hazardous materials shippers. Unlike many of the other safety databases, UNISHIP is not available to the general public. Its primary purpose under Federal Hazardous Materials Transportation Law is to provide DOT administrations with information on past violation histories of hazardous materials offenders for consideration when assessing civil penalties. In addition, information about pending enforcement actions against shippers is also collected and shared, thus allowing each administration to know if another administration is already involved in a pending case. Finally, administrations with active shipper inspection programs can use the information to plan inspections or consolidate enforcement cases across modes. Because the Intermodal Hazardous Materials Programs (IHMP) office of S-3 is in the process of preparing a final Information Resources Management (IRM) procedures document for UNISHIP, no

recommendations have been issued at this time. The IHMP IRM procedures document addresses UNISHIP data file transfer structures and file content issues for improving UNISHIP. It also lists related responsibilities for each Operating Administration (OA), and the Office of the Inspector General (IG). Additional updates to prior years data will be made as required under the final IRM procedures. Currently, the IHMP is working with the Research and Special Programs Administration (RSPA) to develop a schedule for beta-testing bimonthly transfers of UNISHIP data from each of the OAs using the new file content and transfer structures.

Hazardous Materials Information System (HMIS)

The Hazardous Materials Information System (HMIS) consists of six databases that support the mission of the Office of Hazardous Materials Safety (OHMS) in the Research and Special Programs Administration (RSPA). The initial hazardous materials incident reporting system was established in 1971 to meet the requirements of the Hazardous Materials Control Act of 1970. Of the six databases that constitute the HMIS, the only database with a large set of numeric elements with statistical properties is the Hazardous Materials Incident Reporting System (HMIRS).

When an unintentional release of a hazardous material occurs, during transit, loading/unloading, or temporary storage, Title 49 CFR 171.15 and 171.16 requires the transporting carrier to report the incident. Carriers must also notify the National Response Center immediately by telephone, and file an incident report within 30 days,

when any one of the following events occurs:

- one or more major transportation arteries or facilities are closed or shut down for one hour or more,
- the operational flight plan or routine of an aircraft is altered,
- an evacuation occurs lasting one or more hours,
- estimated carrier and/or property damage exceeds \$50,000, or
- a person is killed or hospitalized.

The HMIRS identifies the mode of transportation involved, the name of the reporting carrier, shipment information, the hazardous materials involved, the consequences of the incident, reporter information, and the nature of packaging, as well as the factors contributing to packaging failure. On average, carriers reported about 14,500 incidents per year during 1995 to 1997. The average number of incidents increased to about 17,200 records per year after intrastate reporting started in October 1998.

The assessment conducted by BTS points out positive qualities as well as potential problems within the HMIS system. This data quality assessment is currently under management review.

Airline Passenger Origin-Destination Survey

The Airline Passenger Origin-Destination Survey tracks passengers' use of the commercial air traffic system. It collects information on passenger origins, destinations, and routings. The Civil Aeronautics Board launched the first airline passenger survey in 1947, based on passenger reservations. The reporting basis changed from reservations to tickets in 1968.

After the Civil Aeronautics Board was terminated on December 31, 1984, the Origin-Destination Survey continued as a ticket-based survey under DOT's Research and Special Projects Administration. Since 1995, the Bureau of Transportation Statistics, Office of Airline Information (OAI), has conducted the survey by authorization of Title 14 CFR 241, 19-7.

The Airline Passenger Origin-Destination Survey relies on a 10-percent sample of tickets from large certificated air carriers conducting scheduled passenger services. About 12 million passenger tickets were sampled during 2000. Except for international data (itineraries including non-U.S. points), OAI releases all data from the Airline Passenger Origin-Destination Survey to the public.

Selected findings of the data quality assessment:

- Documentation for the Origin-Destination Survey is weak, thereby hindering the use of the data. The survey also lacks a source and accuracy statement to inform users of the limitations of the data.
- Although tickets are sampled continuously, air carriers report data for the Airline Passenger Origin-Destination Survey 45 days after the end of the quarter, and OAI then takes 75 days to prepare the data.
- Long-term data timeliness and quality gains can be realized if computerized reservation systems can be adapted as the basis for the Origin-Destination Survey. OAI would need to thoroughly test the feasibility and accuracy of such an approach.
- OAI should consider not releasing summary estimates for markets

where the sample size is below a certain minimum, e.g., 200 to 500 cases per quarter. Estimates of variance should be developed for any summary estimates published.

National Transit Database System – Safety & Security Module

The Federal Transit Administration (FTA) has responsibility for the National Transit Database (NTD) system, which is authorized by Title 49 U.S.C. Section 5335(a) and (a)(2). The NTD provides information describing the U.S. transit system with respect to investment, expenditures, operations, and performance. The assessment document provides background information on the NTD, but the actual assessment pertains only to the Safety and Security module.

BTS was given the unique opportunity to assess the pilot for the recent revision of the Safety and Security module. The Safety and Security module is being revised at the direction of the U.S. House and Senate Committees on Appropriations as specified in the Reports to the U.S. Department of Transportation (DOT) FY 2000 Appropriations Act. Recommendations have been made by BTS in the assessment document regarding the collection tool used to gather safety and security information. Current computer technology allows breaking the extensive information required on the Safety and Security Forms into a series of simple questions with appropriate response categories. BTS is recommending that a series of screening questions be used to lead the respondent through the new process, omitting items irrelevant to the mode for which incident data are being reported. It is recommended that response categories be linked to follow-up items for

additional detailed data. In this way, the respondent would only see questions and response categories relevant to their situation. An individual reporting a transit incident should be able to click on a button to view descriptive information pertaining to the questions asked on a particular screen. This can eliminate the need for repetitive references to hardcopy manuals while preparing the report. Built-in controls can prevent the respondent from entering data not of a specific type or outside a prespecified range. The use of such techniques will address certain legislative concerns by:

- Reducing the margin of error – with skip patterns in the instrument, the individuals reporting incident data need not be presented with a list of items and categories, some of which may not apply to them. Individuals reporting incident data for a transit agency that does not operate rail modes, for example, may not be familiar with the terminology and conditions relevant to rail modes.
- Reducing burden of data reporting – by presenting questions and categories specific to the mode involved, the time required for incident reporting is reduced.

The collection of more detailed information on safety and security incidents effectively addresses the congressional mandate to identify common causal factors involved in transit incidents, as specified in the Reports to the U.S. Department of Transportation FY 2000 Appropriation Act. Changing the reporting requirements from annually to monthly or quarterly (depending on the size of the transit authority) will greatly improve the timeliness of the safety and security information on transit incidents.

Because much of the historical data files are not available in text (ascii, or csv) format from NTD website, importing the data into statistical packages, e.g., SAS, proved to be quite difficult and time consuming. BTS recommends that FTA make the NTD data more accessible to users by ensuring that data can be easily and quickly read into statistical packages, like SAS, as well as into user-friendly packages like Excel.

National Aviation Safety Data Analysis Center (NASDAC) Data System

The Federal Aviation Administration's NASDAC data system is a growing, integrated metadata repository. Its User Committee is continuously providing input for enhancements. At present, the NASDAC contains 27 data systems imported from various source databases. There are a number of highly desirable features built into the operational facets of the repository and the number of users is growing. The NASDAC data system currently provides various resources to the aviation community, including:

- a centralized repository of aviation safety databases;
- a library of aviation safety studies and reference materials;
- local and wide area network access;

- Internet, Intranet, and Extranet access;
- data access, analysis, and retrieval software; and
- on-site technical and analytical support personnel.

The NASDAC Data Quality Assessment is being done in stages. This first stage assessment includes a total system overview of NASDAC. This is necessary in order to learn about the data repository, prior to the data quality assessment of individual NASDAC source databases.

Given that the NASDAC warehouses a significant number of various aviation safety databases, the NASDAC User Committee recommended that an automated data quality assessment capability (which could be applied to each source database) should be incorporated into the system. This automated data quality assessment capability is an integral component of NASDAC's Advanced Data Architecture (ADA). The ADA is in the development stage and will be implemented in early 2002. Since NASDAC's data quality assessment and data quality reporting capabilities are still in development stage and are not yet installed in the repository, the recommendations will be deferred until a review of this system component is conducted.

Attachment Data Quality Assessment Template

This is a worksheet for the information gathered and the assessments prepared during a detailed data quality assessment. The information gathered in Sections A through G serves as background material for the Assessment Report, which would consist of an Introduction followed by Sections H and I.

A. Background
1. Name of data system:
2. Sponsoring agency:
3. Legal authority: <i>Legislation, regulations</i>
4. When initiated:
5. Original purpose of data system:
6. Target population: <i>Events/objects/businesses/persons/etc. of interest and rationale for choosing</i>
7. History of data system: <i>Significant changes in purpose, data uses, collection strategies, etc.</i>
8. Future plans: <i>Have any? How are plans formulated?</i>

B. Frames and Sampling (if applicable)
1. Frame: <i>Minimum values for eligibility, sources, update procedures (source? how often updated? how current?), coverage of target population</i>
2. Sample design procedures: <i>Description of sampling technique, stratification/clustering, sample allocation, sample weighting (include post-stratification/benchmarking/calibration), variance estimation, redrawing/rotating (how often?)</i>
3. Sample size: <i>Size of frame, total number selected, number per stage if multistage</i>
4. Documentation: <i>Topics covered, intended audience</i>

C. Data Collection
1. Reporting requirement: <i>Mandatory/voluntary, how enforced</i>
2. Mode of data collection:
3. Frequency of data collection: <i>Periodic (annual/monthly/etc.), irregular/on-demand, e.g., whenever a particular event occurs</i>
4. Geographic coverage: <i>Scope, detail</i>
5. Associated data collection forms and instructions; <i>How are form(s) developed, when and why last changed, pretesting/usability testing</i>
6. Form/instrument: <i>Reference period, summary of content (section by section), due date for completion of form, when data considered usable for reporting purposes, clarity of layout and instructions</i>
7. Number of reports per reporting period:
8. Actual/typical reporter: <i>Number per form, characteristics, knowledge of subject, quality control</i>
9. Amount of effort for reporter/data collector to complete form: <i>Time, research</i>
10. Reporter feedback: <i>Difficulty with form, definitions, availability of information, etc., burden (time, research)</i>
11. Documentation: <i>Topics covered, intended audience</i>

D. Data Preparation
1. Who prepares:
2. Editing: <i>Types of edits, how are error messages dealt with, verification procedures</i>
3. Late/missing reports: <i>Follow-up procedures, rate (and how calculate)</i>
4. Adjustment/imputation for late/missing reports: <i>Procedures, impact on estimates</i>
5. Missing items in reports: <i>Follow-up procedures, rate for key items</i>
6. Any imputation for missing items in reports: <i>Which items, procedures, impact on estimates</i>
7. Changes and updates: <i>Procedures, report files archived?</i>
8. ITDB preparation: <i>Changes made, reasons for changes, impact on estimates</i>
9. Documentation: <i>Topics covered, intended audience</i>

E. Data Dissemination
1. Intended audience: <i>DOT (which part?), Congress, State/local governments, industry/trade associations, researchers, etc.</i>
2. Other major uses (enforcement, etc.):
3. Confidentiality/privacy concerns and protections:
4. Reports and publications: <i>Name, date of release (relative to end of reporting period), particular target audience, format(s) released (hardcopy, online, CD, etc.), how label/identify revisions, description of data limitations included?</i>
5. Analysis: <i>Estimation procedures, statistical comparisons, seasonal/cyclical adjustment</i>
6. Tabular and graphical presentation:
7. Release of data: <i>What information released, what format, available to whom</i>

F. Sponsor Evaluation
1. Coverage of target population:
2. Validation of data:
3. Data quality/limitations of data: <i>Sources and accuracy stated, sampling error, nonsampling error</i>
4. User feedback: <i>Who are actual users, how well are needs met, how is feedback solicited, performance measures collected</i>
5. Prior reviews:

G. Data Quality Staff Data Analysis
1. Ease of access and use
2. Documentation sufficiency, accuracy: <i>Variable names, values, etc.</i>
3. Blank data elements:
4. Overuse of text fields:
5. Coding/classification problems: <i>Mutually exclusive and exhaustive, systematic, overuse of "other", "NEC", etc.</i>
6. Duplicate records:
7. Outliers:
8. Inconsistencies among items:
9. Ability to reproduce published/official estimates:
10. Relationship to other data: <i>Within data system over time, ability to relate to external data systems (e.g., standard definitions, codes), estimates, duplication between systems</i>
11. Anything else that looks strange:
12. Source of data used in DQ staff analysis: <i>Name of file, location, date acquired/accessed, version, etc.</i>

H. Assessment
1. Relevance and Completeness: <i>User needs and data gaps, coverage of major issues, user involvement mechanisms, program review and monitoring policies [cf. Sections A, E, F]</i>
2. Quality: <i>Design of data collection meet objectives, how carefully implemented, assessments of accuracy provided, quantification of accuracy and deficiencies [cf. Sections B, C, D, E, G]</i>
3. Timeliness: <i>Delay between reference date and time information available, delay between time information available and time needed to be useful [cf. Sections A, C, E, F]</i>
4. Comparability: <i>Ability to combine with other information, cross-modal consistency in concepts and definitions, consistency with non-DOT concepts and definitions [cf. B, C, G]</i>
5. Utility: <i>Ease of obtaining information, suitability of format for users, availability of supplementary information/metadata needed to use data correctly, documentation, and its interpretability, statements describing limitations of the data [cf. Sections A, C, D, E, F, G]</i>

I. Recommendations and Suggestions for Data Quality Improvements
1. “Tactical” (correcting any errors found during review):
2. “Strategic” (improving procedures): <i>Easy (low-lying fruit), hard (e.g., need additional resources)</i>
3. Continuing: <i>Follow-up on implementation of recommendations, development of standards</i>

J. References